GLARE: GENERALIZABLE LENS-AGNOSTIC RELATIVE-RAY ENCODING FOR 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 3D reconstruction models learn transferable geometric priors, but changes in camera projection can limit their applicability across lens configurations. To address this challenge, we present **GLARE** (**G**eneralizable **L**ens-**A**gnostic **R**elative-Ray **E**ncoding), an image-only adaptation framework for diverse camera projections. Our approach uses image-inferred viewing directions to condition feature comparison and information exchange within a pretrained reconstruction transformer. Inference requires only RGB images, without calibration, camera poses, or lens-family labels. Extensive evaluations across diverse datasets demonstrate that our method consistently establishes new state-of-the-art performance, yielding substantial reductions in 3D reconstruction and depth estimation errors while significantly improving camera-pose accuracy. These results support image-inferred observation geometry as an effective interface for adapting pretrained multi-view reconstruction across camera domains.

1 INTRODUCTION

Feed-forward 3D foundation models predict scene geometry and camera pose from image collections in a single pass (Wang et al., 2024; 2025; 2026). Their transferable priors are promising for robotics, autonomous driving, world models, and AR/VR, yet are learned predominantly from perspective imagery. These applications involve diverse camera systems, whose projections change the mapping from image location to viewing direction, whereas image-grid coordinates alone do not expose that observation geometry. Retraining a foundation model for every camera family requires substantial computation and suitable multi-view data, motivating our central question: ***Can a perspective-pretrained 3D foundation model adapt to diverse camera or lens configurations using only RGB images?***

Recent work extends reconstruction beyond perspective cameras through domain- or task-specific models, explicit camera conditioning, and post-hoc alignment (Duan et al., 2026; Li et al., 2026; Guo et al., 2026; Yuan et al., 2026; Jung et al., 2026; Guruprasad et al., 2026). In parallel, lightweight camera conditioning has adapted pretrained transformers without relearning their full visual prior (Zhang et al., 2026). Together these developments suggest intervening in the latent interactions of a reconstruction model rather than addressing projection only at its input or output.

We present **GLARE**, **G**eneralizable **L**ens-**A**gnostic **R**elative-Ray **E**ncoding for 3D Reconstruction. A continuous image-conditioned ray field associates each patch token with a representative viewing direction. These rays define orthogonal feature actions whose pairwise composition conditions both comparison and value transport in a parallel multi-view adapter. This design makes observation geometry available where features from different views interact. Calibration and poses provide training supervision; inference uses RGB images alone.

We evaluate reconstruction, depth, and camera pose in fisheye, ERP, and perspective images. The complete evaluations show improvements over other models in diverse camera image datasets, while qualitative comparisons illustrate more complete room layouts and more coherent reconstructed surfaces.

Our contributions are threefold:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1. We introduce GLARE, a novel framework for lens-agnostic feed-forward 3D reconstruction that infers continuous camera-local ray fields directly from RGB images, completely bypassing the need for known camera calibration, poses, or domain labels at inference.

2. We propose a lightweight geometry-mediated parallel adapter that embeds 3D observation geometry into latent multi-view feature interactions. By formulating ray-conditioned orthogonal actions for both feature comparison and value transport, our method effectively adapts perspective-pretrained foundation models to diverse projections while preserving their frozen visual encoders and native aggregators.

3. We demonstrate robust and consistent improvements in 3D reconstruction, depth, and camera pose estimation across diverse camera domains. Compared to prior baselines, our approach significantly reduces geometry inaccuracies, achieving up to a 60.7% reduction in reconstruction completion errors and a 43.8% decrease in depth estimation errors, while notably boosting stringent camera pose accuracy (AUC@5°) by up to 47.7 percentage points.

2 Related Work

2.1 Feed-forward 3D Reconstruction across Camera Projections

General-purpose feed-forward 3D reconstruction methods directly infer 3D scene representations, entirely bypassing the need for test-time bundle optimization (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025; 2026; Keetha et al., 2026). Transformer scaling and heterogeneous data yield transferable priors, but their positional and interaction structure is predominantly learned from perspective cameras. Camera-general approaches therefore use domain-specific adaptation, camera-family conditioning, ray-aware representations, or learned calibration. For instance, UniK3D estimates spherical-harmonic ray fields for monocular camera-general prediction (Piccinelli et al., 2025), while Fisheye3R, PanoVGGT, and Argus specialize reconstruction architectures to fisheye or panoramic imagery (Duan et al., 2026; Guo et al., 2026; Li et al., 2026), and VGGT360 targets panoramic depth (Yuan et al., 2026). However, these domain-specific networks are not designed for cross-lens generalization. To bridge diverse domains, Wid3R conditions its aggregator with an explicit camera-model token (Jung et al., 2026), whereas CAM3R infers rays in a two-view network and merges pairwise predictions by post-hoc alignment (Guruprasad et al., 2026). Both approaches impose practical constraints: explicit tokens demand prior knowledge of the camera type, while post-hoc alignments complicate the feed-forward inference pipeline. We instead estimate camera-local rays from images and seamlessly embed them inside latent multi-view aggregation without camera tokens or post-hoc alignment at inference. Our focus is adapting the latent interactions of a pretrained multi-view reconstruction model while retaining its robust point and pose prediction interface.

2.2 Camera-conditioned Transformer Interactions

Camera rays, intrinsics, extrinsics, and depth can effectively ground transformer interactions in 3D. PRoPE encodes known camera frustums as relative positional information (Li et al., 2025), and RayRoPE constructs projective ray encodings for attention over posed views (Wu et al., 2026). Beyond 3D reconstruction, UCPE injects known pose, intrinsics, and lens distortion into video-diffusion transformers through a lightweight parallel attention adapter (Zhang et al., 2026); CRePE extends this conditioning with depth-aware distributions along projected paths (Jin et al., 2026). A critical limitation across these approaches is their strict reliance on explicitly supplied camera parameters or ground-truth poses during inference. Inspired by the parameter-efficient parallel conditioning strategy from video generation, we combine native spatial encoding with ray-dependent orthogonal actions in a reconstruction adapter. Crucially, our setting lifts the requirement for provided metadata by dynamically inferring the observation field directly from images. In our design, predicted camera-local directions condition feature comparison and value transport, while camera pose is jointly estimated by the reconstruction network. This seamlessly connects camera-conditioned attention to RGB-only multi-view reconstruction without requiring any supplied camera metadata.

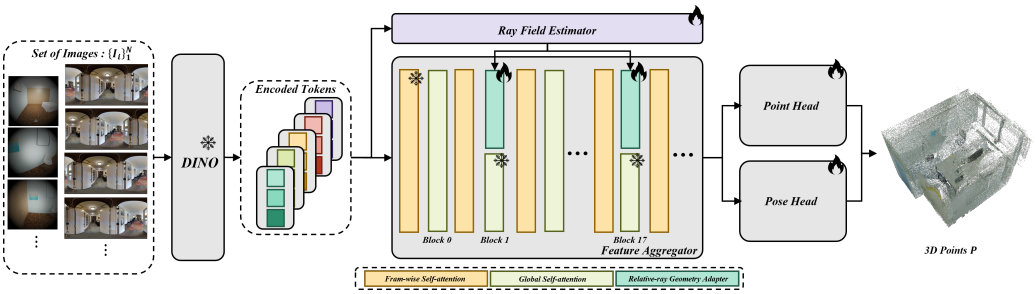

Figure 1: Overview of our reconstruction pipeline. An image-conditioned Ray-Field Estimator provides camera-local viewing directions to a lightweight geometry-conditioned adaptation path within the pretrained multi-view reconstruction model. The network jointly predicts camera-local scene geometry and camera-to-world poses from RGB images.

3 METHOD

Our model connects image-conditioned ray estimation to multi-view feature aggregation and joint point and pose prediction (Figure 1). We first define the reconstruction task, then describe how inferred rays enter the attention computation and how the outputs are supervised.

3.1 PROBLEM FORMULATION

Given an unordered RGB image set $\mathcal{I} = \{I_i\}_{i=1}^N$ observing the same scene, f_Θ predicts a camera-local point map and a camera-to-world pose for each view,

$$f_\Theta\big(\{I_i\}_{i=1}^N\big) = \big\{\big(\mathbf{X}_i^{\text{local}}, T_i\big)\big\}_{i=1}^N, \qquad T_i = (R_i, t_i) \in \mathrm{SE}(3). \tag{1}$$

$I_i \in \mathbb{R}^{H\times W\times 3}$ is an RGB image, $\mathbf{X}_i^{\text{local}} \in \mathbb{R}^{H\times W\times 3}$ is expressed in camera-i coordinates, and $R_i \in \mathrm{SO}(3), t_i \in \mathbb{R}^3$ map it to a common reconstruction frame. We jointly infer geometry and pose across camera projections from RGB images, without calibration, camera parameters, or camera-family labels.

Geometric motivation. For a static scene point $\mathbf{P}_\ell^*$ observed at $\mathbf{u}_{i\ell}$, a central camera obeys (Ramalingam et al., 2006)

$$\mathbf{P}_\ell^* = t_i^* + \varrho_{i\ell} R_i^* \mathbf{b}_i(\mathbf{u}_{i\ell}), \qquad \mathbf{b}_i(\mathbf{u}_{i\ell}) \in \mathbb{S}^2, \quad \varrho_{i\ell} > 0. \tag{2}$$

The ideal bearing $\mathbf{b}_i$ specifies camera-local observation direction, while pose and radial distance determine its scene realization. Cross-view bearings constrain motion and structure (Hartley & Zisserman, 2004, Chapter 9); Appendix A gives the coordinate derivation. We therefore expose predicted directions $\mathbf{r}_i$ during feature aggregation without imposing Eq. 2 as an output constraint.

3.2 IMAGE-CONDITIONED RAY FIELD ESTIMATION

For image I_i, the Ray Field Estimator (RFE) predicts a continuous unit field $\mathbf{r}_i : \Omega_i = [0,1]^2 \to \mathbb{S}^2$. Following UniK3D's spherical-harmonic parameterization (SH) (Piccinelli et al., 2025), a Transformer encoder processes frozen ViT features and predicts image-level domain parameters ϑ_i and 15 3D coefficient vectors $C_i \in \mathbb{R}^{15\times 3}$. The field is

$$\widetilde{\mathbf{r}}_i(\mathbf{u}) = C_i^\top \mathbf{Y}_{1:3}(\mathbf{s}(\mathbf{u};\vartheta_i)), \qquad \mathbf{r}_i(\mathbf{u}) = \frac{\widetilde{\mathbf{r}}_i(\mathbf{u})}{\|\widetilde{\mathbf{r}}_i(\mathbf{u})\|_2}. \tag{3}$$

Here $\mathbf{s}(\mathbf{u};\vartheta_i) \in \mathbb{S}^2$ is an auxiliary spherical query and $\mathbf{Y}_{1:3} \in \mathbb{R}^{15}$ contains real Cartesian SH functions. The decoded $\mathbf{r}_i(\mathbf{u})$ is the camera-local ray. We evaluate this image-level field at patch centers; the same samples supervise RFE and condition aggregation. Appendix B specifies details.

3.3 RAY-CONDITIONED MULTI-VIEW AGGREGATION

The Relative-Ray Geometry Adapter (RRGA) uses the inferred field to condition interactions among patch tokens. For token $a = (i, p)$, let $\mathbf{r}_a$ be the ray evaluated at its normalized patch center (Appendix A). For a unit ray, define the feature-space action

$$A(\mathbf{r}) = \mathbf{r}\mathbf{r}^\top + [\mathbf{r}]_\times \in \mathrm{SO}(3), \tag{4}$$

where $[\mathbf{r}]_\times x = \mathbf{r} \times x$. This is a fixed-$\pi/2$ rotation about $\mathbf{r}$, used as an orthogonal feature action rather than a physical camera rotation.

To integrate both native 2D image context and 3D observation geometry, we partition each 64-dimensional attention head into two distinct subspaces. Specifically, we retain 28 channels for native image-grid rotary cues, denoted as $D_{\mathrm{native},a} \in \mathrm{SO}(28)$, and allocate the remaining 36 channels to twelve 3-channel ray blocks. The joint token-wise operator $D_a \in \mathrm{SO}(64)$ and its pairwise composition U_{ab} are defined via a direct sum:

$$D_a = D_{\mathrm{native},a} \oplus \bigoplus_{m=1}^{12} A(\mathbf{r}_a) \in \mathrm{SO}(64), \quad U_{ab} = D_a^\top D_b. \tag{5}$$

Through this formulation, 2D position and image-inferred 3D geometry jointly condition every head. While both heads within the adapter share the same operator D_a, they maintain independent learned projections. Importantly, the relative-ray encoding U_{ab} operates strictly as a feature-space transformation rather than a physical camera-frame rotation.

Inspired by UCPE's parallel conditioning design (Zhang et al., 2026), our adapter receives the same pre-block features as a frozen global block and attends over patch tokens from all views. Independent affine projections produce 64-dimensional Q, K, and V for head h. Comparison uses

$$\ell_{ab}^h = \frac{(q_a^h)^\top U_{ab} k_b^h}{\sqrt{64}}, \qquad \alpha_{ab}^h = \mathrm{softmax}_b(\ell_{ab}^h), \tag{6}$$

and the same pairwise operator transports values with a query-specific back-transform,

$$y_a^h = D_a^\top \sum_b \alpha_{ab}^h D_b v_b^h = \sum_b \alpha_{ab}^h U_{ab} v_b^h. \tag{7}$$

The resulting branch geometrically conditions, rather than replaces, pretrained multi-view reasoning. Its two heads are concatenated and added in parallel:

$$x_{\mathrm{out}} = F_{\mathrm{frozen}}(x) + P_{\mathrm{out}}\,\mathrm{Concat}_h(y^h). \tag{8}$$

Here F_{frozen} is the native global block and P_{out} is the learned output projection. Both paths receive the same pre-block features x. Zero-initializing P_{out} makes the initial adapter residual zero, while native attention, positional encoding, and register paths remain unchanged. Additional implementation details and operator properties are provided in Appendices C and D.

3.4 POINT AND POSE PREDICTION

The pretrained point and pose branches are initialized from the π^3 checkpoint and jointly optimized. We replace its perspective-oriented point mapping with unrestricted camera-local XYZ, permitting signed coordinates including $z < 0$. The pose branch predicts $T_i = (R_i, t_i)$, giving

$$\mathbf{X}_i^{\mathrm{world}}(\mathbf{u}) = R_i \mathbf{X}_i^{\mathrm{local}}(\mathbf{u}) + t_i. \tag{9}$$

Rays condition latent interactions but do not hard-constrain final point directions, and predicted poses are not fed back into the adapter. Points and translations are expressed in the normalized units defined in Section 3.5.

3.5 TRAINING OBJECTIVES

A sampled training group g contains $N_g \geq 2$ views. We normalize GT local points and translations by a group-shared scale γ_g, defined as the median radius of all valid GT local points and lower-bounded by 10^{-4}. Predictions directly regress these normalized targets without prediction-side scale fitting. Point and translation residuals use norm-based Smooth-L1 with $\beta = 0.1$, while rotation is supervised by $1 - \cos\theta$. Exact normalization, masking, and reduction rules are provided in Appendix E.

Local geometry. For each view, the predicted camera-local point map is directly supervised against the normalized GT geometry:

$$\mathcal{L}_{\text{local}} = \operatorname*{mean}_{g,i} \frac{1}{|\mathcal{M}_{g,i}|} \sum_{\mathbf{u} \in \mathcal{M}_{g,i}} \rho_\beta \left(\mathbf{X}^{\text{local}}_{g,i}(\mathbf{u}) - \mathbf{Y}^{*}_{g,i}(\mathbf{u}) \right), \tag{10}$$

where $\mathcal{M}_{g,i}$ denotes the valid point support and $\mathbf{Y}^{*}_{g,i} = \mathbf{X}^{*\text{local}}_{g,i} / \gamma_g$.

Relative camera pose. For predicted camera-to-world poses $\widehat{T}_{g,i}$, we form the relative transform $\widehat{T}_{g,ij} = \widehat{T}^{-1}_{g,i} \widehat{T}_{g,j} = (\widehat{R}_{g,ij}, \widehat{t}_{g,ij})$, which maps camera-$j$ coordinates into camera-i coordinates. For every ordered pair $i \neq j$,

$$\mathcal{L}_{\text{pose}} = \operatorname*{mean}_{g, i \neq j} \left[1 - \text{clamp} \left(\frac{\text{tr}(\widehat{R}^{\top}_{g,ij} R^{*}_{g,ij}) - 1}{2}, -1, 1 \right) + \rho_\beta \left(\widehat{t}_{g,ij} - \frac{t^{*}_{g,ij}}{\gamma_g} \right) \right]. \tag{11}$$

Common-frame reconstruction. To couple local geometry and camera motion, each source-view point map is transformed into every reference-camera frame. For reference k and source i,

$$\widehat{\mathbf{Y}}_{g,k \leftarrow i} = \widehat{R}_{g,ki} \mathbf{X}^{\text{local}}_{g,i} + \widehat{t}_{g,ki}, \qquad \mathbf{Y}^{*}_{g,k \leftarrow i} = R^{*}_{g,ki} \mathbf{Y}^{*}_{g,i} + \frac{t^{*}_{g,ki}}{\gamma_g}. \tag{12}$$

We then minimize

$$\mathcal{L}_{\text{world}} = \operatorname*{mean}_{g,k,i} \frac{1}{|\mathcal{M}_{g,i}|} \sum_{\mathbf{u} \in \mathcal{M}_{g,i}} \rho_\beta \left(\widehat{\mathbf{Y}}_{g,k \leftarrow i}(\mathbf{u}) - \mathbf{Y}^{*}_{g,k \leftarrow i}(\mathbf{u}) \right), \tag{13}$$

for all reference–source pairs, including $i = k$. This term directly supervises point–pose composition in a common 3D coordinate frame. It does not perform prediction-side scale alignment or introduce an explicit 2D reprojection, epipolar, or bundle-adjustment objective.

Ray-field supervision. The RFE is additionally supervised by calibrated camera-local unit rays. For the valid patch-center support $\mathcal{Q}_{g,i}$,

$$\mathcal{L}_{\text{RFE}} = \operatorname*{mean}_{g,i} \frac{1}{|\mathcal{Q}_{g,i}|} \sum_{\mathbf{u} \in \mathcal{Q}_{g,i}} \left\| \mathbf{r}_{g,i}(\mathbf{u}) - \mathbf{r}^{*}_{g,i}(\mathbf{u}) \right\|_2^2. \tag{14}$$

Ray supervision constrains the observation geometry available to the representation, whereas the three reconstruction objectives supervise local scene geometry, relative camera motion, and their common-frame composition.

The complete objective is

$$\mathcal{L} = \lambda_{\text{local}} \mathcal{L}_{\text{local}} + \lambda_{\text{pose}} \mathcal{L}_{\text{pose}} + \lambda_{\text{world}} \mathcal{L}_{\text{world}} + \lambda_{\text{RFE}} \mathcal{L}_{\text{RFE}}, \tag{15}$$

with

$$(\lambda_{\text{local}}, \lambda_{\text{pose}}, \lambda_{\text{world}}, \lambda_{\text{RFE}}) = (1, 1.5, 1.5, 0.5). \tag{16}$$

Valid pixels are averaged within each view, actual views or view pairs are averaged within each group, and groups are weighted equally. Full target construction, robust penalties, masks, reductions, and the conditional common-frame consistency analysis are given in Appendix E.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We train on three data sources—ScanNet++, ASE v2, and RealseeDB and reserve Stanford2D3D and FIORD for zero-shot evaluation. Appendix G details their roles and geometric interface.

Baselines. We compare with VGGT, π^3, fine-tuned π^3 (π^{3*}), and Wid3R.

Metrics. To evaluate 3D reconstruction quality, we report point-cloud accuracy, completion, and normal consistency to assess both spatial proximity and surface orientation. For radial depth estimation, we measure AbsRel, RMSE, and the standard threshold accuracy (δ_1). Camera pose estimation is assessed using Relative Rotation Accuracy (RRA), Relative Translation Accuracy (RTA), and joint pose Area Under the Curve (AUC), all rigorously evaluated at a stringent 5° error threshold.

Training. Starting from pretrained π^3, we jointly optimize RFE, the adapters, and the point/pose branches; the image encoder and native aggregator remain frozen. We use 448×224 inputs, randomly sample $N \in [2, 24]$ views per group, and optimize with AdamW at a learning rate of 10^{-4} Appendix F gives the initialization, schedule, and optimizer details.

4.2 3D RECONSTRUCTION

Across datasets spanning heterogeneous camera projections, GLARE shows a consistent trend toward more balanced reconstruction quality, improving point accuracy and coverage on most benchmarks while strengthening surface-normal agreement across the evaluated domains. This pattern suggests that the gains are not limited to point-wise proximity, but also extend to the geometric coherence of the recovered surfaces. Cases where distance errors and normal consistency rank methods differently further indicate that these metrics capture complementary aspects of reconstruction quality. Overall, GLARE maintains stronger reconstruction fidelity across camera domains rather than benefiting from a single benchmark-specific setting; the full per-dataset results are reported in Table 1.

Figure 2 shows that the qualitative differences are most pronounced for non-perspective inputs. Under fisheye and ERP projections, several baselines produce fragmented or spatially inconsistent geometry, while GLARE more consistently preserves the global room structure and surface coherence. The gap is less pronounced on perspective scenes, where all methods recover broadly similar layouts. This pattern is consistent with Table 1, where the largest gains are observed in reconstruction coverage and surface agreement under heterogeneous camera projections.

Table 1: 3D reconstruction across five datasets. Acc.: accuracy error; Comp.: completion error; N.C.: normal consistency. Bold and underline mark the best and second-best entries.

	Realsee Real			**Realsee Synthetic**			**ASE**			**ScanNet++**			**Stanford2D3D**		
Method	**Acc. ↓**	**Comp. ↓**	**N.C. ↑**	**Acc. ↓**	**Comp. ↓**	**N.C. ↑**	**Acc. ↓**	**Comp. ↓**	**N.C. ↑**	**Acc. ↓**	**Comp. ↓**	**N.C. ↑**	**Acc. ↓**	**Comp. ↓**	**N.C. ↑**
VGGT	0.422	1.706	0.521	0.464	1.223	0.574	0.446	1.673	0.543	0.155	0.089	0.731	0.520	2.561	0.526
π^3	0.413	1.316	0.539	0.486	1.019	<u>0.576</u>	<u>0.188</u>	0.267	0.776	0.093	0.054	0.794	0.542	2.075	0.581
π^{3*}	0.277	0.344	0.535	<u>0.254</u>	<u>0.230</u>	0.527	**0.160**	**0.080**	0.716	<u>0.077</u>	<u>0.044</u>	0.746	<u>0.292</u>	<u>0.245</u>	0.613
Wid3R	<u>0.157</u>	<u>0.185</u>	<u>0.783</u>	0.427	0.369	0.559	0.286	0.322	<u>0.787</u>	0.085	0.046	<u>0.817</u>	0.369	0.284	<u>0.749</u>
Ours	**0.075**	**0.073**	**0.868**	**0.037**	**0.031**	**0.892**	0.235	<u>0.247</u>	**0.797**	**0.035**	**0.029**	**0.823**	**0.288**	**0.244**	**0.777**

4.3 DEPTH ESTIMATION

The depth results show a clear reduction in both relative and absolute radial-depth error, with GLARE achieving the lowest reported AbsRel and RMSE on all five benchmarks. This consistency across perspective, fisheye, and ERP datasets indicates that the improvement is not tied to a particular camera domain. Threshold accuracy, however, does not always follow the same ordering: on ASE, the lowest average errors and the highest δ_1 are achieved by different methods, indicating that mean depth deviation and tolerance-based accuracy capture complementary properties of the

error distribution. Overall, GLARE exhibits stronger radial-depth fidelity across camera domains rather than a benchmark-specific gain; full quantitative comparisons are provided in Table 2.

Table 2: Depth estimation across five evaluation datasets. AbsRel and RMSE are lower-is-better; δ_1 is higher-is-better. Bold and underline mark the best and second-best entries.

	Realsee Real			Realsee Synthetic			ASE			ScanNet++			Stanford2D3D		
Method	**AbsRel ↓**	**RMSE ↓**	**δ_1 ↑**	**AbsRel ↓**	**RMSE ↓**	**δ_1 ↑**	**AbsRel ↓**	**RMSE ↓**	**δ_1 ↑**	**AbsRel ↓**	**RMSE ↓**	**δ_1 ↑**	**AbsRel ↓**	**RMSE ↓**	**δ_1 ↑**
VGGT	0.527	1.268	0.301	0.372	0.822	0.447	0.664	1.738	0.279	0.154	0.514	0.798	0.374	1.473	0.473
π^3	0.382	0.978	0.429	0.270	0.547	0.597	0.132	0.394	0.944	0.089	0.387	0.947	0.274	1.098	0.602
π^{3*}	0.184	<u>0.619</u>	0.717	0.150	0.479	0.811	<u>0.045</u>	0.228	**0.990**	0.072	0.357	0.960	0.130	0.791	0.821
Wid3R	<u>0.102</u>	0.768	<u>0.911</u>	<u>0.071</u>	<u>0.350</u>	<u>0.957</u>	0.045	<u>0.218</u>	0.977	<u>0.066</u>	<u>0.274</u>	<u>0.973</u>	<u>0.103</u>	<u>0.633</u>	**0.924**
Ours	**0.057**	**0.500**	**0.962**	**0.031**	**0.179**	**0.987**	**0.042**	**0.198**	<u>0.977</u>	**0.053**	**0.164**	**0.977**	**0.100**	**0.520**	<u>0.924</u>

4.4 Camera Pose Estimation

The camera-pose results show that GLARE improves both rotation and translation-direction estimation across most camera domains, rather than benefiting only one component of relative motion. Its stronger RRA and RTA are also reflected in the joint pose AUC, indicating that the gains persist when rotation and translation are evaluated jointly. Generalization is nevertheless not uniform across held-out domains: GLARE retains the strongest reported results on FIORD, while π^{3*} remains better on Stanford2D3D. This pattern indicates stronger camera-pose estimation overall, while also highlighting residual dataset-specific variation in cross-domain transfer; detailed results are given in Table 3.

Table 3: Camera-pose estimation at a common 5° threshold across six datasets. All metrics are percentages (higher is better). Bold and underline mark the best and second-best entries; π^{3*} denotes the fine-tuned model.

	Realsee Real			Realsee Synthetic			ASE		
Method	RRA@5° ↑	RTA@5° ↑	AUC@5° ↑	RRA@5° ↑	RTA@5° ↑	AUC@5° ↑	RRA@5° ↑	RTA@5° ↑	AUC@5° ↑
VGGT	4.45	2.60	0.00	20.50	5.34	0.08	0.31	0.13	0.00
π^3	2.28	3.74	0.14	24.47	2.59	0.33	0.34	0.10	0.00
π^{3*}	36.37	41.46	7.56	<u>95.28</u>	<u>70.06</u>	<u>26.27</u>	<u>7.70</u>	<u>14.77</u>	<u>0.59</u>
Wid3R	<u>45.60</u>	<u>51.88</u>	<u>17.41</u>	63.40	0.98	0.16	0.61	0.15	0.00
Ours	**97.85**	**96.93**	**65.11**	**100.00**	**98.08**	**84.66**	**75.82**	**72.29**	**40.08**

	ScanNet++			FIORD			Stanford2D3D		
Method	RRA@5° ↑	RTA@5° ↑	AUC@5° ↑	RRA@5° ↑	RTA@5° ↑	AUC@5° ↑	RRA@5° ↑	RTA@5° ↑	AUC@5° ↑
VGGT	27.12	16.29	2.18	5.62	5.72	0.59	4.68	3.71	0.06
π^3	50.64	24.97	5.40	10.01	6.04	0.41	4.34	4.70	0.06
π^{3*}	<u>91.73</u>	52.78	14.92	<u>13.82</u>	<u>12.32</u>	<u>2.51</u>	**87.32**	**86.85**	**59.70**
Wid3R	85.50	<u>70.85</u>	<u>37.27</u>	1.18	1.13	0.00	76.26	75.65	36.09
Ours	**96.37**	**88.84**	**58.91**	**42.64**	**29.47**	**6.73**	<u>81.68</u>	<u>83.26</u>	<u>48.94</u>

5 Conclusion

We presented GLARE, an RGB-only approach to adapting pretrained multi-view reconstruction across camera projections. Its central idea is to expose image-inferred viewing geometry within latent feature interactions. Relative-ray conditioning acts on both feature comparison and value transport, allowing the model to jointly learn scene geometry and camera pose while preserving the frozen encoder and native aggregator. This provides a practical way to reuse pretrained geometric priors with diverse lenses without supplying calibration, poses, or camera-family labels at inference.

The full evaluations show lower reconstruction accuracy error and completion error, higher normal consistency, and lower depth AbsRel than Wid3R on all five dense benchmarks. On Realsee Real, accuracy and completion errors decrease by 52.2% and 60.7%, respectively, alongside a 43.8% AbsRel reduction and a 47.7 percentage points gain in pose AUC@5°. The 59.0% accuracy-error reduction on ScanNet++ and improved reconstruction on held-out Stanford2D3D extend the evidence across camera and dataset domains. Qualitative examples complement these measurements with more complete layouts and coherent surfaces. The lower strict-threshold pose accuracy on

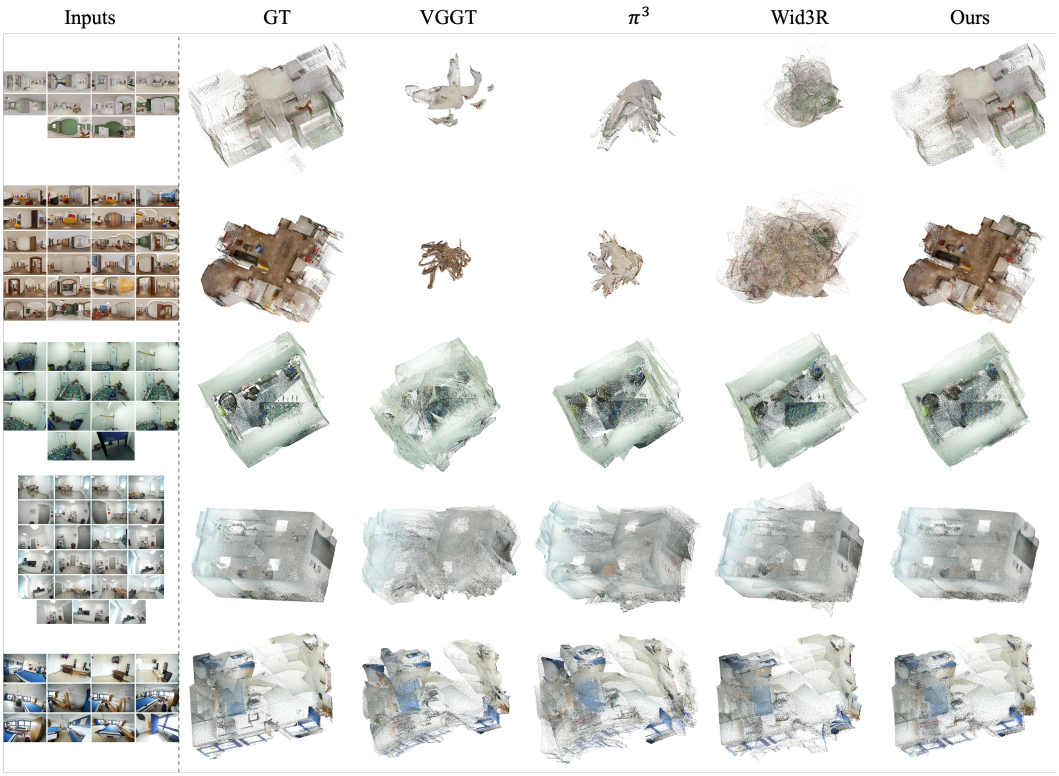

Figure 2: Qualitative 3D reconstruction across camera domains. Columns show the input-view grid, GT, VGGT, π^3, Wid3R, and our model.

FIORD and the remaining gap to fine-tuned π^3 on Stanford2D3D identify limits in held-out pose generalization. Together, the findings support conditioning pretrained feature interactions on inferred observation geometry as an effective strategy for camera-general reconstruction.

Reproducibility Statement

Section 3 specifies the overall model formulation and training objectives. Appendix A defines the coordinate and geometric conventions used throughout the model. Appendices B and C provide the implementation details of the Ray Field Estimator and Relative-Ray Geometry Adapter, respectively, including the ray-field parameterization, attention dimensions, and feature-space operators. Appendix D provides the mathematical properties and scope of the ray-conditioned operator. Appendix E specifies the target construction, scale normalization, robust penalties, validity supports, loss reductions, and objective weights. Appendix F reports the initialization strategy, optimizer hyperparameters, learning-rate schedule, data-source sampling strategy, and trainable and frozen components. Appendix G describes the roles of the training and evaluation datasets, the geometric supervision interface, validity handling, and preprocessing used to construct point and ray supervision. Section 4.1 defines the evaluation tasks and metrics used in the reported experiments. Source code, configurations, dataset-processing scripts, and evaluation scripts will be released with the final version to facilitate replication of the reported results.

AI Use Statement

Generative AI tools assisted with manuscript restructuring, consistency checks between the manuscript and implementation, and software-development support. The authors reviewed the resulting text, equations, code, and citations and take responsibility for the complete content.

REFERENCES

Ruxiao Duan, Erin Hong, Dongxu Zhao, Eric Turner, Alex Wong, and Yunwen Zhou. Fisheye3R: Adapting unified 3d feed-forward foundation models to fisheye lenses. In *European Conference on Computer Vision*, pp. 1–19, 2026.

Yijing Guo, Mengjun Chao, Luo Wang, Tianyang Zhao, Haizhao Dai, Yingliang Zhang, Jingyi Yu, and Yujiao Shi. PanoVGGT: Feed-forward 3d reconstruction from panoramic imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36444–36453, 2026.

Namitha Guruprasad, Abhay Yadav, Cheng Peng, and Rama Chellappa. CAM3R: Camera-agnostic model for 3d reconstruction. In *European Conference on Computer Vision*, pp. 227–246, 2026.

Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. URL `https://www.robots.ox.ac.uk/~vgg/hzbook/`.

Seonghyun Jin, Youngmin Kim, Sunwoo Park, and Jong Chul Ye. CRePE: Curved ray expectation positional encoding for unified-camera-controlled video generation, 2026.

Dongki Jung, Jaehoon Choi, Adil Qureshi, Somi Jeong, Dinesh Manocha, and Suyong Yeon. Wid3R: Wide field-of-view 3d reconstruction via camera model conditioning. In *European Conference on Computer Vision*, 2026.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3d reconstruction. In *International Conference on 3D Vision*, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R, 2024.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding, 2025.

Xi Li, Linyuan Li, Yan Wu, Tong Rao, Kai Zhang, Xinchen Hui, and Cihui Pan. Argus: Metric panoramic 3d reconstruction for indoor scenes. In *European Conference on Computer Vision*, pp. 608–627, 2026.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3d estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1028–1039, 2025.

Srikumar Ramalingam, Suresh K. Lodha, and Peter Sturm. A generic structure-from-motion framework. *Computer Vision and Image Understanding*, 103:218–228, 2006. doi: 10.1016/j.cviu.2006.06.006.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment—a modern synthesis. In Bill Triggs, Andrew Zisserman, and Richard Szeliski (eds.), *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pp. 298–372. Springer, 2000. URL `https://www.cs.jhu.edu/~misha/ReadingSeminar/Papers/Triggs00.pdf`.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. In *International Conference on Learning Representations*, 2026.

Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. RayRoPE: Projective ray positional encoding for multi-view attention, 2026.

Jiayi Yuan, Haobo Jiang, De Wen Soh, and Na Zhao. VGGT-360: Geometry-consistent zero-shot panoramic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19874–19883, 2026.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38027–38037, 2026.

A Notation and coordinate conventions

RFE coordinates and native patch positions obey separate conventions. At the default raster, RFE processes a 16×32 patch grid, giving 512 center queries per view in row-major patch order. For token $a = (i, p)$ at patch row j_p and column k_p, define

$$\mathbf{u}_a = \left(\frac{14(k_p + 1/2)}{W}, \frac{14(j_p + 1/2)}{H} \right), \qquad \mathbf{r}_a = \mathbf{r}_i(\mathbf{u}_a). \tag{A1}$$

At $W \times H = 448 \times 224$, the first and last centers are $(7/448, 7/224)$ and $(441/448, 217/224)$. The corresponding 0-based pixel-index coordinate is $(14(k_p + 1/2) - 1/2, 14(j_p + 1/2) - 1/2)$, and the `align_corners=False` sampling grid is $(2u_a - 1, 2v_a - 1)$. No additional half-pixel offset or nearest-integer lookup is applied. The endpoint-normalized coordinate $\mathbf{c}_p$ used inside the RFE encoder and the native π^3 positions (π_a^x, π_a^y) remain separate conventions from the field-evaluation coordinate $\mathbf{u}_a$.

Bearing observations and relative geometry. This paragraph concerns ideal observations, not additional inputs or operations in our model. Let $\mathbf{b}_{i\ell} = \mathbf{b}_i(\mathbf{u}_{i\ell})$ in Eq. 2. Under the camera-to-world convention, define

$$T_{ij}^* = (T_i^*)^{-1} T_j^*, \qquad R_{ij}^* = (R_i^*)^\top R_j^*, \qquad t_{ij}^* = (R_i^*)^\top (t_j^* - t_i^*). \tag{A2}$$

Thus T_{ij}^* maps camera-j coordinates to camera-i coordinates. Equating observations of the same static point and expressing them in camera i gives

$$\varrho_{i\ell} \mathbf{b}_{i\ell} = \varrho_{j\ell} R_{ij}^* \mathbf{b}_{j\ell} + t_{ij}^*. \tag{A3}$$

Define $[t]_\times v = t \times v$. Left multiplication by $\mathbf{b}_{i\ell}^\top [t_{ij}^*]_\times$ eliminates the first and translation terms because

$$\mathbf{b}_{i\ell}^\top [t_{ij}^*]_\times \mathbf{b}_{i\ell} = 0, \qquad [t_{ij}^*]_\times t_{ij}^* = 0. \tag{A4}$$

Since $\varrho_{j\ell} > 0$, this yields the calibrated essential relation (Hartley & Zisserman, 2004, Chapter 9),

$$\mathbf{b}_{i\ell}^\top E_{ij}^* \mathbf{b}_{j\ell} = 0, \qquad E_{ij}^* = [t_{ij}^*]_\times R_{ij}^*. \tag{A5}$$

Corresponding bearings thus constrain unknown relative geometry. Under nondegenerate observations, central-camera reconstruction retains a global similarity ambiguity; a bearing field alone does not determine pose or absolute scale. Projection rays provide an observation representation across camera models (Ramalingam et al., 2006). Pixels themselves also contain no depth: their pixel-to-bearing mapping is a calibration-dependent re-expression of the observations. Bundle adjustment

Table A1: Coordinate and supervision conventions used throughout our method.

Symbol / convention	Meaning
W, H	Input width and height; the default raster is 448×224 ($W \times H$).
(i, p, h)	View, patch, and RRGA-head indices.
(u, v)	Continuous normalized edge coordinates; a pixel center is $((x + 0.5)/W, (y + 0.5)/H)$.
$\mathbf{c}_p$	Endpoint-normalized RFE patch-grid coordinate on $[-1, 1]^2$.
(π_a^x, π_a^y)	Native π^3 patch-grid position in horizontal/vertical order.
$\mathbf{s}_i$	Auxiliary unit direction used only to evaluate the SH basis.
$\mathbf{r}_i$	Predicted camera-local unit ray in the x-right, y-down, z-forward convention.
C_i	Fifteen 3D SH coefficient vectors, stored as a 15×3 matrix.
$\mathbf{u}_a, \mathbf{r}_a$	Normalized patch-center coordinate and its predicted camera-local unit ray.
$A(\mathbf{r})$	Direct $\mathrm{SO}(3)$ feature action parameterized by unit ray $\mathbf{r}$.
$D_{\mathrm{native},a}$	Reduced native image-grid rotary operator in $\mathrm{SO}(28)$.
$D_{\mathrm{geo},a}$	Head-shared center-ray geometry operator in $\mathrm{SO}(36)$.
D_a	Native/geometry direct sum in $\mathrm{SO}(64)$, shared by both heads.
U_{ab}	Pairwise feature operator $D_a^\top D_b$.
$I_d, O(d)$	$d \times d$ identity and orthogonal group; $^\top$ denotes transpose.
T_i	Camera-to-world transform; $T_{ki} = T_k^{-1} T_i$ maps view-i coordinates to view k.
g, k	Sampled multi-view group and reference-view index in the loss.
γ_g	GT median-radius normalization computed once for group g.
$\mathcal{M}_{g,i}$	Valid point-map pixels; $\mathcal{Q}_{g,i}$ is valid sampled-ray support.

jointly adjusts shared points and camera parameters to explain fixed image observations (Triggs et al., 2000). These classical relations motivate the use of viewing directions; they are neither a solver nor a constraint implemented by our feature-space operator.

B Ray Field Estimator details

For 448×224 input views, the frozen ViT encodes 512 patch tokens per view. Two T-Enc blocks are applied sequentially with eight heads, expansion four, zero dropout, and layer scale 10^{-4} (Piccinelli et al., 2025). In each block, LayerNorm is applied to the feature input before the Q/K/V projections; the same coordinate embedding is then added to projected Q and K, but not directly to V. This additive position handling belongs to RFE and is distinct from the full-D Q/K/V action in RRGA. Mean pooling is applied only after both blocks, followed by LayerNorm and a biased $256 \to 48$ readout. No CLS token, camera label, or cross-view feature enters RFE.

The readout contains three raw domain scalars and 15 three-dimensional SH coefficient vectors. For raw scalars $(\eta_i^f, \eta_i^x, \eta_i^y)$, the implemented domain map is

$$\phi_i^h = 2\pi\sigma(\eta_i^f - 1.1), \qquad \phi_i^v = \phi_i^h H/W, \qquad \bar{c}_i^x = \sigma(\eta_i^x), \qquad \bar{c}_i^y = \sigma(\eta_i^y). \tag{A6}$$

For $\mathbf{u} = (u, v)$, longitude and latitude are $\lambda_i = (u - \bar{c}_i^x)\phi_i^h$ and $\varphi_i = (v - \bar{c}_i^y)\phi_i^v$. The auxiliary SH direction is

$$\mathbf{s}(\mathbf{u}; \vartheta_i) = [\cos\varphi_i \sin\lambda_i \quad -\sin\varphi_i \quad \cos\varphi_i \cos\lambda_i]^\top. \tag{A7}$$

The real Cartesian SH basis excludes the degree-zero constant, leaving $3+5+7 = 15$ functions. The 45 stored coefficient scalars are organized as 15 row vectors in $C_i \in \mathbb{R}^{15\times3}$; they are not 45 distinct SH basis functions. Only the degree-one coefficient bias is initialized to the $(s_x, -s_y, s_z)$ field, and all coefficients remain trainable. Other affine weights use the module's standard initialization, while the final readout weight uses a small initialization.

The field parameters are estimated once per image and evaluated at all patch centers. The resulting sampled-ray tensor in $[B, N, P, 3]$ form is reused by both heads, all adapted RRGA blocks, and the

RFE loss. It contains evaluations of the image-level field rather than a new field parameterization; the same evaluator remains applicable at arbitrary coordinates.

C Relative-ray geometry adapter implementation

Each RRGA maps $1024 \to 128$ independently for Q, K, and V and splits the result into two 64-dimensional heads. Every head contains 28 native and 36 geometry channels:

$$q_a^h = P_Q^h \,\mathrm{LN}(x_a), \qquad k_a^h = P_K^h \,\mathrm{LN}(x_a), \qquad v_a^h = P_V^h \,\mathrm{LN}(x_a). \tag{A8}$$

The corresponding unary actions are

$$\widetilde{q}_a^h = D_a q_a^h, \qquad \widetilde{k}_a^h = D_a k_a^h, \qquad \widetilde{v}_a^h = D_a v_a^h. \tag{A9}$$

For a unit ray, the implementation applies

$$A(\mathbf{r}) = \mathbf{r}\mathbf{r}^\top + [\mathbf{r}]_\times, \qquad A(\mathbf{r})x = \mathbf{r}(\mathbf{r}^\top x) + \mathbf{r} \times x, \qquad A(\mathbf{r})^\top x = \mathbf{r}(\mathbf{r}^\top x) - \mathbf{r} \times x. \tag{A10}$$

For $\mathbf{r} = (r_x, r_y, r_z)^\top$, the skew matrix is $[\mathbf{r}]_\times = \begin{bmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{bmatrix}$. Writing the geometry slice as $x_{\mathrm{geo}} = [(x^{[1]})^\top, \ldots, (x^{[12]})^\top]^\top$ with $x^{[m]} \in \mathbb{R}^3$, the shared operator is

$$D_{\mathrm{geo},a} = \bigoplus_{m=1}^{12} A(\mathbf{r}_a), \quad D_{\mathrm{geo},a} x_{\mathrm{geo}} = [(A(\mathbf{r}_a)x^{[1]})^\top, \ldots, (A(\mathbf{r}_a)x^{[12]})^\top]^\top. \tag{A11}$$

These channels hold native horizontal, native vertical, and geometry features. Non-finite inputs or rays with norm at most 10^{-8} are rejected before normalization; valid vectors are divided by their actual norm, and transformed features are cast back to the attention dtype. Concatenated head outputs are projected $128 \to 1024$. Expanding the pairwise composition gives

$$U_{ab} = D_a^\top D_b = (D_{\mathrm{native},a}^\top D_{\mathrm{native},b}) \oplus \bigoplus_{m=1}^{12} \left[A(\mathbf{r}_a)^\top A(\mathbf{r}_b) \right]. \tag{A12}$$

The final three channels are interpreted directly as unrestricted camera-local coordinates,

$$\mathbf{X}_i^{\mathrm{local}}(\mathbf{u}) = \mathrm{PointBranch}(\mathbf{F}_i)(\mathbf{u}) \in \mathbb{R}^3. \tag{A13}$$

Here $\mathbf{F}_i$ denotes the aggregated features for view i supplied to the point branch.

D Properties of the ray-conditioned orthogonal action

Orthogonality. For a unit ray, let $P = \mathbf{r}\mathbf{r}^\top$ and $S = [\mathbf{r}]_\times$. The identities $P^2 = P$, $PS = SP = 0$, and $S^2 = P - I$ give

$$A(\mathbf{r})^\top A(\mathbf{r}) = (P - S)(P + S) = P - S^2 = I. \tag{A14}$$

Moreover, $A(\mathbf{r})$ is a rotation by $\pi/2$ about $\mathbf{r}$, so $\det A(\mathbf{r}) = 1$ and $A(\mathbf{r}) \in \mathrm{SO}(3)$. Also, $A(\mathbf{r}) - A(\mathbf{r})^\top = 2[\mathbf{r}]_\times$; hence the individual unary mapping retains the ray's direction and sign, but does not encode a camera pose.

Norm preservation. Eq. A14 implies $\|A(\mathbf{r})x\|_2 = \|x\|_2$. Direct sums of the native and geometry blocks therefore satisfy

$$D_a^\top D_a = I_{64}, \qquad U_{ab} = D_a^\top D_b \in \mathrm{SO}(64), \qquad U_{ab}^\top U_{ab} = I_{64}. \tag{A15}$$

Cauchy–Schwarz then yields

$$|\ell_{ab}^h| \le \frac{\|q_a^h\|_2 \|k_b^h\|_2}{\sqrt{d_h}}, \qquad d_h = 64. \tag{A16}$$

This controls norm amplification by the deterministic operator for fixed projected features; it does not establish network stability, reconstruction robustness, or ray-estimation robustness.

Coordinate covariance. For $Q \in \mathrm{SO}(3)$, $(Q\mathbf{r})(Q\mathbf{r})^\top = Q\mathbf{r}\mathbf{r}^\top Q^\top$ and $[Q\mathbf{r}]_\times = Q[\mathbf{r}]_\times Q^\top$. Hence

$$A(Q\mathbf{r}) = QA(\mathbf{r})Q^\top. \tag{A17}$$

The ray-to-action mapping thus transforms consistently under a change of 3D ray coordinates; this is not a claim of whole-network rotation equivariance.

Ray error and operator error. For unit rays $\mathbf{r}$ and $\mathbf{s}$, let $\mu = \mathbf{r}^\top \mathbf{s}$, $M = \mathbf{r}\mathbf{r}^\top - \mathbf{s}\mathbf{s}^\top$, and $N = [\mathbf{r} - \mathbf{s}]_\times$, so $A(\mathbf{r}) - A(\mathbf{s}) = M + N$. Since M is symmetric and N is skew-symmetric, their Frobenius inner product vanishes. Therefore

$$\begin{aligned} \|M\|_F^2 &= 2(1-\mu^2), \\ \|N\|_F^2 &= 2\|\mathbf{r}-\mathbf{s}\|_2^2 = 4(1-\mu), \\ \|A(\mathbf{r}) - A(\mathbf{s})\|_F^2 &= 2(1-\mu^2) + 4(1-\mu) = (3+\mu)\|\mathbf{r}-\mathbf{s}\|_2^2, \\ 2\|\mathbf{r}-\mathbf{s}\|_2^2 &\le \|A(\mathbf{r}) - A(\mathbf{s})\|_F^2 \le 4\|\mathbf{r}-\mathbf{s}\|_2^2. \end{aligned} \tag{A18}$$

Thus squared chordal RFE error controls deterministic operator error up to bounded constants. It does not bound attention outputs, correspondence error, or reconstruction error.

Physical-geometry boundary. For bearings in cameras i and j, physical world-direction comparison requires

$$\mathbf{r}_a^\top R_i^\top R_j \mathbf{r}_b. \tag{A19}$$

The unknown relative rotation $R_i^\top R_j$ is not an RRGA input. Consequently, $A(\mathbf{r}_a)^\top A(\mathbf{r}_b)$ is a relative feature-space composition based on camera-local viewing directions, not a physical relative camera transform.

E Complete Training Objectives and Reductions

This section specifies the target construction, robust penalties, validity supports, and reduction rules used by the training objectives in Sec. 3.5. The definitions below correspond directly to Eqs. 10–15; we therefore avoid repeating the individual loss expressions.

E.1 Group-wise scale normalization

A training group g contains $N_g \ge 2$ actual views. Let $\mathcal{M}_{g,i}$ denote the nonempty set of valid GT point pixels for view i, and define the valid group support as

$$\mathcal{M}_g = \{(i, \mathbf{u}) \ : \ \mathbf{u} \in \mathcal{M}_{g,i}\}. \tag{A20}$$

Unstarred quantities denote predictions and starred quantities denote ground truth.

We use a single GT-derived scale for all views in the group,

$$\gamma_g = \max\left(\underset{(i,\mathbf{u})\in\mathcal{M}_g}{\operatorname{median}} \left\|\mathbf{X}^{*\mathrm{local}}_{g,i}(\mathbf{u})\right\|_2, 10^{-4}\right), \tag{A21}$$

and define the normalized camera-local point target as

$$\mathbf{Y}^*_{g,i}(\mathbf{u}) = \frac{\mathbf{X}^{*\mathrm{local}}_{g,i}(\mathbf{u})}{\gamma_g}. \tag{A22}$$

The same γ_g divides all GT translations associated with group g. Predicted local points and translations directly regress these normalized targets; no prediction-dependent scale is estimated or fitted during either training or loss evaluation. Rotations and unit rays are unaffected by this normalization.

Because γ_g is shared across the group rather than computed independently per view, relative scene scale among the views is preserved within each training sample.

E.2 Robust point and translation penalty

All 3D point and translation residuals use Smooth-L1 applied to the Euclidean residual norm rather than independently to Cartesian components. For $\mathbf{\Delta} \in \mathbb{R}^3$,

$$\rho_\beta(\mathbf{\Delta}) = \frac{\mathrm{Huber}_\beta\left(\|\mathbf{\Delta}\|_2\right)}{\beta} = \begin{cases} \dfrac{\|\mathbf{\Delta}\|_2^2}{2\beta}, & \|\mathbf{\Delta}\|_2 < \beta, \\ \|\mathbf{\Delta}\|_2 - \dfrac{\beta}{2}, & \text{otherwise}, \end{cases} \qquad \beta = 0.1. \tag{A23}$$

Thus, the penalty behaves quadratically near zero and linearly for larger geometric residuals, while preserving rotational symmetry in 3D residual space.

E.3 Relative-pose convention

Both predicted and GT poses are represented as camera-to-world transformations. For an ordered pair (i, j), the relative transforms used in Eq. 11 are

$$\widehat{T}_{g,ij} = \widehat{T}^{-1}_{g,i}\widehat{T}_{g,j} = \left(\widehat{R}_{g,ij}, \widehat{t}_{g,ij}\right), \qquad T^*_{g,ij} = (T^*_{g,i})^{-1}T^*_{g,j} = \left(R^*_{g,ij}, t^*_{g,ij}\right). \tag{A24}$$

Accordingly, $T_{g,ij}$ maps coordinates from camera j to camera i.

The relative rotation error is evaluated as

$$1 - \cos\theta_{g,ij} = 1 - \mathrm{clamp}\left(\frac{\mathrm{tr}\left(\widehat{R}^\top_{g,ij}R^*_{g,ij}\right) - 1}{2}, -1, 1\right), \tag{A25}$$

where clamping is applied for numerical stability. Relative translation targets are normalized by the same group scale,

$$\widetilde{t}^*_{g,ij} = \frac{t^*_{g,ij}}{\gamma_g}. \tag{A26}$$

All ordered pairs $i \neq j$ are included in the pose objective.

E.4 Common-frame target construction

The common-frame objective in Eq. 13 uses the same normalized local geometry and normalized translations as the local and pose objectives. For reference camera k and source camera i,

$$\widehat{\mathbf{Y}}_{g,k\leftarrow i}(\mathbf{u}) = \widehat{R}_{g,ki}\mathbf{X}^{\mathrm{local}}_{g,i}(\mathbf{u}) + \widehat{t}_{g,ki}, \tag{A27}$$

and the corresponding GT target is

$$\mathbf{Y}^*_{g,k\leftarrow i}(\mathbf{u}) = R^*_{g,ki}\mathbf{Y}^*_{g,i}(\mathbf{u}) + \frac{t^*_{g,ki}}{\gamma_g}. \tag{A28}$$

The validity support is inherited from the source-view point map: $\mathbf{u} \in \mathcal{M}_{g,i}$. No target correspondence between source and reference images is required to evaluate this objective. All reference–source combinations are included, including $i = k$.

For $i = k$, the relative transform is the identity in the ideal case, so the common-frame target reduces to the normalized camera-local target. For $i \neq k$, the term explicitly couples predicted local geometry and predicted relative motion through 3D composition.

This objective is evaluated directly in normalized 3D camera coordinates. It is not a 2D reprojection loss and does not introduce an additional prediction-side scale alignment.

E.5 RFE TARGET CONSTRUCTION

Let $\mathcal{Q}_{g,i}$ denote the nonempty valid patch-center support used for RFE supervision in Eq. 14. Dense GT rays are first constructed from the calibrated camera model in camera-local coordinates. They are then bilinearly sampled at the patch-center locations and renormalized to unit length:

$$\mathbf{r}^*_{g,i}(\mathbf{u}) \leftarrow \frac{\widetilde{\mathbf{r}}^*_{g,i}(\mathbf{u})}{\left\|\widetilde{\mathbf{r}}^*_{g,i}(\mathbf{u})\right\|_2}, \qquad \mathbf{u} \in \mathcal{Q}_{g,i}. \tag{A29}$$

Ray validity is sampled using nearest-neighbor interpolation so that interpolation itself does not create valid supervision at an invalid spatial location. The validity mask is used only to select supervised targets and is not provided as an input to the forward attention mechanism.

The squared Cartesian ray residual

$$\left\|\mathbf{r}_{g,i}(\mathbf{u}) - \mathbf{r}^*_{g,i}(\mathbf{u})\right\|_2^2 \tag{A30}$$

is summed over the three Cartesian components before reduction. Each valid patch center contributes once, independently of the number of attention heads, feature blocks, or adapted layers that subsequently consume the ray representation.

E.6 HIERARCHICAL GROUP-BALANCED REDUCTION

The notation mean in Sec. 3.5 denotes a hierarchical reduction rather than a single global average over all valid pixels in a minibatch.

For a generic per-view pixelwise loss $\ell_{g,i}(\mathbf{u})$ with valid support $\mathcal{S}_{g,i}$, we first compute

$$\bar{\ell}_{g,i} = \frac{1}{|\mathcal{S}_{g,i}|} \sum_{\mathbf{u}\in\mathcal{S}_{g,i}} \ell_{g,i}(\mathbf{u}), \tag{A31}$$

and then average the actual views in each group,

$$\bar{\ell}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} \bar{\ell}_{g,i}. \tag{A32}$$

Finally, groups are weighted equally:

$$\mathcal{L} = \frac{1}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \bar{\ell}_g, \tag{A33}$$

where $\mathcal{G}$ denotes the groups in the minibatch.

The same principle is applied to objectives defined over view combinations. For the ordered-pair pose objective, all valid $i \neq j$ pairs are first averaged within group g:

$$\bar{\ell}^{\mathrm{pose}}_g = \frac{1}{N_g(N_g-1)} \sum_{i\neq j} \ell^{\mathrm{pose}}_{g,ij}. \tag{A34}$$

For the common-frame objective, all reference–source combinations are averaged within the group:

$$\bar{\ell}_g^{\text{world}} = \frac{1}{N_g^2} \sum_{k=1}^{N_g} \sum_{i=1}^{N_g} \bar{\ell}_{g,k,i}^{\text{world}}. \tag{A35}$$

Groups are then averaged with equal weight.

Consequently, a training group does not receive a larger optimization weight merely because it contains more valid pixels or a larger number of sampled views.

E.7 Conditional common-frame consistency

The common-frame reconstruction objective also provides a useful interpretation of cross-view geometric consistency. Consider a fixed group g and reference camera k. Suppose valid source observations $\mathbf{u}_i$ and $\mathbf{u}_j$ correspond to the same physical scene point ℓ. Under consistent GT geometry,

$$\mathbf{Y}^*_{g,k\leftarrow i}(\mathbf{u}_i) = \mathbf{Y}^*_{g,k\leftarrow j}(\mathbf{u}_j) = \mathbf{Y}^*_{g,k,\ell}. \tag{A36}$$

Define their individual reconstruction residuals as

$$\epsilon_i = \left\| \widehat{\mathbf{Y}}_{g,k\leftarrow i}(\mathbf{u}_i) - \mathbf{Y}^*_{g,k,\ell} \right\|_2, \qquad \epsilon_j = \left\| \widehat{\mathbf{Y}}_{g,k\leftarrow j}(\mathbf{u}_j) - \mathbf{Y}^*_{g,k,\ell} \right\|_2. \tag{A37}$$

The triangle inequality gives

$$\left\| \widehat{\mathbf{Y}}_{g,k\leftarrow i}(\mathbf{u}_i) - \widehat{\mathbf{Y}}_{g,k\leftarrow j}(\mathbf{u}_j) \right\|_2 \le \epsilon_i + \epsilon_j. \tag{A38}$$

Hence, when two observations of the same physical point are individually reconstructed accurately in the same reference-camera frame, their predictions must also be close to one another in that frame. This property follows from supervision against a shared 3D target; correspondence labels are not consumed by the model and no additional pairwise consistency term is required.

This implication is conditional on the individual residuals. A small average value of $\mathcal{L}_{\text{world}}$ does not imply that every physical correspondence has a small reconstruction error.

E.8 Final weighting

The result-generating configuration uses

$$(\lambda_{\text{local}}, \lambda_{\text{pose}}, \lambda_{\text{world}}, \lambda_{\text{RFE}}) = (1, 1.5, 1.5, 0.5), \tag{A39}$$

for the complete objective in Eq. 15.

The local, pose, and common-frame terms supervise scene geometry and camera motion at complementary levels, while the RFE objective supervises the camera-local observation geometry supplied to the representation. None of these objectives introduces an explicit epipolar-correspondence, 2D reprojection, or bundle-adjustment constraint.

F Optimization and training protocol

The image encoder, alternating aggregator, and point/pose modules are initialized from a pretrained π^3 checkpoint. RFE is initialized independently. RRGA Q/K/V projections use independent standard initialization, while only the output projection is zero initialized so that the adapter residual initially vanishes. The result-generating run jointly optimizes RFE, RRGA, and the pretrained point/pose branches from the first update, without a separate pretraining stage or learning-rate warm-up. The ray-conditioned $\mathrm{SO}(3)$ action contains no learned or selected geometry frequency; its axis is the normalized predicted ray and its rotation angle is fixed to $\pi/2$.

All optimizer groups use AdamW with learning rate 10^{-4}, betas $(0.9, 0.95)$, weight decay 0.01, and gradient clipping at norm 1. The 64,000-update run uses cosine decay to 0.01 of the base rate without a learning-rate warm-up. Training uses a common 448×224 raster and random view counts from 2 to 24. One source is selected per update from ASE v2, ScanNet++, Realsee Synthetic, and Realsee

Table A2: Datasets and their roles in the evaluation protocol.

Dataset	Projection domain	Role
ScanNet++	perspective / distorted	training and ID evaluation
ASE v2	fisheye	training and ID evaluation
Realsee Synthetic / Real	ERP	training and ID evaluation
Stanford2D3D	ERP	Zero-shot evaluation only
FIORD	fisheye	Zero-shot evaluation only

Real. Each source has raw configuration weight 1.0, yielding normalized sampling probability 0.25. Multi-view groups share the group-level geometric normalization defined in Section 3.5. The image encoder and native alternating aggregator remain frozen, while autograd through their operations is retained so reconstruction losses can reach RFE and RRGA after the initially zero output projection receives an update. At the exact zero-initialized state, that projection blocks upstream reconstruction gradients; RFE still receives its direct ray-field loss.

G DATA PROTOCOL

All training sources are converted to a common interface comprising RGB images, camera-to-world poses, camera-local point maps, dense calibration ray maps, and separate point/ray validity masks. Projection-native geometry is regenerated after resizing each input to 448×224, including the source-specific crop and rotation convention. These quantities form supervision only; the forward model receives RGB images. After resize/crop/rotation preprocessing, dense calibrated ray maps are sampled at the patch centers. Invalid camera regions are excluded from the corresponding loss, and every sampled training view must retain nonempty support.

The four training sources use the sampling weights and group-level scale convention reported in Appendix F. Stanford2D3D and FIORD remain disjoint from optimization and checkpoint selection. Metrics that require unavailable ground-truth geometry are not reported; in particular, the current FIORD interface does not provide dense reconstruction targets.